

[0008]

[Means for solving problem] The HTML descriptive information extraction display system in the first invention comprises a character string retrieving means for retrieving a keyword character string including a keyword that has been specified by descriptive information described in Hyper Media Description Language, a tag analysis means for analyzing a tag in the descriptive information and detecting a divided tag that has been specified in advance, an information extraction means for extracting a keyword character string from the descriptive information on the basis of the divided tag that has been extracted and forming it into excerpted information described in Hyper Media Description Language, and a display means for displaying the extracted keyword character string.

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平11-134341

(43) 公開日 平成11年(1999) 5月21日

(51) Int.Cl.⁴

識別記号

F I

G 0 6 F 17/30

G 0 6 F 15/401

3 2 0 A

15/40

3 4 0

15/403

3 8 0 E

審査請求 有 請求項の数 7 O L (全 6 頁)

(21) 出願番号 特願平9-292806

(71) 出願人 000004237

日本電気株式会社

東京都港区芝五丁目7番1号

(22) 出願日 平成9年(1997)10月24日

(72) 発明者 新井 良和

東京都港区芝五丁目7番1号 日本電気株式会社内

(74) 代理人 弁理士 京本 直樹 (外 2 名)

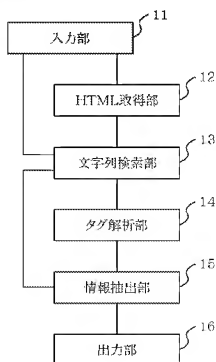
(54) 【発明の名称】 ハイパーメディア記述言語による記述情報の抜粋表示シ

ステム

(57) 【要約】

【課題】 HTML記述情報の中から必要な部分を自動的に抜粋して一つに纏めて表示することができ、本当に必要な情報の掲載されたページを迅速にかつ、効率的に探索することができるHTML記述情報抜粋表示システムの提供。

【解決手段】 入力部 11 から入力されたURLによりHTML取得部 12 が取得したHTMLファイルを文字列検索部 13 によりキーワードを含む文字列を検索し、この文字列に含まれている分割タグをタグ解析部 14 により見つけてこれによりキーワードを含む文字列を情報抽出部 15 により抽出して出力部 16 に表示する。



【特許請求の範囲】

【請求項1】 ハイパーメディア記述言語により記述された記述情報から指定されたキーワードを含むキーワード文字列を検索する文字列検索手段と、前記記述情報にあるタグを解析し予め指定された分割タグを検出するタグ解析手段と、前記検出された分割タグに基づいて前記キーワード文字列を前記記述情報から抽出しハイパーメディア記述言語による抜粋情報に整形する情報抽出手段と、前記抽出されたキーワード文字列を表示する表示手段とを含むことを特徴とするハイパーメディア記述言語による記述情報の抜粋表示システム。

【請求項2】 文字列検索手段は検索に際しては先ず記述情報からタグを削除してキーワード文字列を検索し検索後に前記削除したタグを復元することを特徴とする請求項1記載のハイパーメディア記述言語による記述情報の抜粋表示システム。

【請求項3】 情報抽出手段はキーワードに先行する最も近い先行指定分割タグとキーワードに後続する最も近い後続指定分割タグとを検索して前記先行指定分割タグと後続指定分割タグとを含むその間の文字列を抽出することを特徴とする請求項1記載のハイパーメディア記述言語による記述情報の抜粋表示システム。

【請求項4】 情報抽出手段は抽出したキーワードを含む文字列ブロックが複数あるときには文字列区切りタグにより1つの文字列ブロックにすることを特徴とする請求項3記載のハイパーメディア記述言語による記述情報の抜粋表示システム。

【請求項5】 情報抽出手段は抽出した文字列ブロックの先頭に終了タグがある場合および前記文字列ブロックの終端に開始タグがある場合にはこれを削除し、これ以外で前記文字列ブロック中に先行する開始タグに対応する後続する終了タグが存在する場合以外には対応する不足タグを追加することを特徴とする請求項3記載のハイパーメディア記述言語による記述情報の抜粋表示システム。

【請求項6】 情報抽出手段はハイパーメディア記述言語によることを示すタグと、情報のヘッダ部を示すタグと、情報の本文を示すタグとにより抜粋情報に整形することを特徴とする請求項3記載のハイパーメディア記述言語による記述情報の抜粋表示システム。

【請求項7】 ハイパーメディア記述言語による記述情報であるホームページ情報をネットワークを介して取得し前記ホームページ情報の指定されたキーワードを含む抜粋情報を作成して表示することを特徴とする請求項1記載のハイパーメディア記述言語による記述情報の抜粋表示システム。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明はハイパーメディア記述言語（以下HTMLと略称する：HyperText

Markup Language）による記述情報の抜粋表示システムに関し、特にWWW（World Wide Web）におけるHTMLで記述されたホームページ情報の中から与えられたキーワードに関連した部分を抜粋表示するHTML記述情報抜粋表示システムに関する。

【0002】

【従来の技術】最近急速に普及してきたインターネットにおいては、WWWと呼ばれる情報検索システムが構築されており、このWWWによれば、種々の情報やサービスがHTMLと呼ばれる言語によって記述されたホームページにより提供されるようになっている。

【0003】現在、ホームページの数は、莫大な数となっており、その中からユーザが自身によって所望のものを見つけ出すのは困難である。そこで、WWWでは、例えばキーワードなどを入力すると、そのキーワードを含むホームページを検索するような種々の検索サービスが提供されている。

【0004】さらには特開平9-171513号公報記載の技術では、ユーザが、自身が所望するホームページが開設されたことを、容易に認識することもできるようになっている。

【0005】そしてこのキーワード検索を行って対応するホームページのURL（Uniform Resource Locator、一種のアドレス）を得て、これを用いて、WWWブラウザによりホームページを表示してユーザが所望の情報を得ている。

【0006】

【発明が解決しようとする課題】上述した従来のホームページの表示システム、すなわち、HTML記述情報表示システムは、ホームページ、すなわち、HTML記述情報の全文を表示しており、上記のキーワード検索で多数のURLが得られた場合にはその全てのホームページをWWWブラウザで表示して、本当に必要な情報の掲載されたページを目で探さなければならないという欠点を有している。

【0007】本発明の目的は、ホームページ、すなわち、HTML記述情報の中から必要と考えられる部分を自動的に抜粋して一つに纏めて表示できるHTML記述情報抜粋表示システムを提供することにある。

【0008】

【課題を解決するための手段】第1の発明のHTML記述情報抜粋表示システムは、ハイパーメディア記述言語により記述された記述情報から指定されたキーワードを含むキーワード文字列を検索する文字列検索手段と、前記記述情報にあるタグを解析し予め指定された分割タグを検出するタグ解析手段と、前記検出された分割タグに基づいて前記キーワード文字列を前記記述情報から抽出しハイパーメディア記述言語による抜粋情報に整形する情報抽出手段と、前記抽出されたキーワード文字列を表

示する表示手段とを含んで構成されている。

【0009】第2の発明のHTML記述情報抜粋表示システムは、第1の発明のHTML記述情報抜粋表示システムにおいて、文字列検索手段は検索に際しては先ず記述情報からタグを削除してキーワード文字列を検索し検索後に前記削除したタグを復元することを特徴としている。

【0010】第3の発明のHTML記述情報抜粋表示システムは、第1の発明のHTML記述情報抜粋表示システムにおいて、情報抽出手段はキーワードに先行する最も近い先行指定分割タグとキーワードに後続する最も近い後続指定分割タグとを検索して前記先行指定分割タグと後続指定分割タグとを含むその間の文字列を抽出することを特徴としている。

【0011】第4の発明のHTML記述情報抜粋表示システムは、第3の発明のHTML記述情報抜粋表示システムにおいて、情報抽出手段は抽出したキーワードを含む文字列ブロックが複数あるときには文字列区切りタグにより1つの文字列ブロックにすることを特徴としている。

【0012】第5の発明のHTML記述情報抜粋表示システムは、第3の発明のHTML記述情報抜粋表示システムにおいて、情報抽出手段は抽出した文字列ブロックの先頭に終了タグがある場合および前記文字列ブロックの終端に開始タグがある場合にはこれを削除し、これ以外で前記文字列ブロック中に先行する開始タグに対応する後続する終了タグが存在する場合以外は対応する不足タグを追加することを特徴としている。

【0013】第6の発明のHTML記述情報抜粋表示システムは、第3の発明のHTML記述情報抜粋表示システムにおいて、情報抽出手段はハイパーメディア記述言語によることを示すタグと、情報のヘッダ部を示すタグと、情報の本文を示すタグとにより抜粋情報に整形することを特徴としている。

【0014】第7の発明のHTML記述情報抜粋表示システムは、第1の発明のHTML記述情報抜粋表示システムにおいて、ハイパーメディア記述言語による記述情報であるホームページ情報をネットワークを介して取得し前記ホームページ情報の指定されたキーワードを含む抜粋情報を作成して表示することを特徴としている。

【0015】

【発明の実施の形態】次に、本発明の実施の形態について図面を参照して説明する。

【0016】図1は本発明のHTML記述情報抜粋表示システムの一実施の形態を示すブロック図である。

【0017】本実施の形態のHTML記述情報抜粋表示システムは、図1に示すように、URLとキーワードとを入力する入力部11と、入力部11から受け取り対応するHTMLファイルをWWWサーバから取得するHT

ML取得部12と、入力部11からキーワードを受け取りHTML取得部12からHTMLファイルを受けとりHTMLファイルに対しキーワード文字列の検索を行なう文字列検索部13と、文字列検索部13からキーワード文字列の検索の終了したHTMLファイルを受けとりそのタグを検索しキーワード文字列周辺で文書構造の区切りに使用されることが多いタグを探し出すタグ解析部14とキーワードを含みタグ解析部14で探し出されたタグで囲まれた部分を抜き出し文書ブロックとしさらにその文書ブロックが複数ある場合はこれを一つにまとめて表示可能なHTMLファイル形式にする情報抽出部15と、情報抽出部15で得られるHTMLファイルを表示する出力部16とから構成されている。

【0018】これらにより構成される本実施の形態のHTML記述情報抜粋表示システムは、図2に示すキーボード等の入力装置21と、情報を表示するディスプレイ22と、入力部11、HTML取得部12、文字列検索部13、タグ解析部14、情報抽出部15、出力部16等の処理を行なうコンピュータ23とにより実現される。

【0019】図3は本実施の形態のHTML記述情報抜粋表示システムの動作を示す流れ図である。図1～図3を参照して本実施の形態のHTML記述情報抜粋表示システムの動作を説明する。

【0020】まず、入力装置21から与えられたURLとキーワードは、コンピュータ23に実装された入力部11によってURLはHTML取得部12へ、キーワードは文字列検索部13へ渡される(ステップ302)。

【0021】URLを受け取ったHTML取得部12は対応するHTMLファイルをWWWサーバから取得する(ステップ303)。

【0022】次に、文字列検索部13は入力部11からキーワード、HTML取得部12からHTMLファイルを受け取り、HTMLファイルに対しキーワード検索を行なう(ステップ304)。この際HTMLのタグとコメント部分は検索を行なう前に一旦取り除き検索を行ない、検索終了後に元あった場所にタグとコメント部分を戻すようにした方がよい。これは、タグとコメントはHTMLファイルを参照する際、眼には触れない部分であり、キーワードの検索の対象にならないし、また、それがあることにより検索に誤動作を与えるためである。

【0023】キーワードに当てはまる部分がなかった場合は(ステップ305のN枝)、その旨を出力部16に伝え表示する(ステップ312)。

【0024】キーワードに当てはまる部分があった場合は(ステップ305のY枝)、そのキーワードの位置を記憶して置く(ステップ306)。

【0025】次にタグ解析部14はキーワード検索の終了したHTMLファイルを受け取りその文書の区切りに使われるタグを検索する(ステップ307)。そして見

つけだされたタグを元にして情報抽出部15はキーワードを含む部分を抜き出し文書ブロックにする(ステップ308)。文書ブロックの作成処理は図4を用いて後述する(ステップ308)。

【0026】作成した文書ブロックが複数あるかチェックし(ステップ309)、複数あるときには(ステップ309のY枝)、改行タグ
、水平線タグ<HR>等の区切り用タグを挟んでつなぎあわせ一つの文書ブロックにする(ステップ310)。この例は図5を用いて後述する。

【0027】次に、HTMLの必須タグ(図6(b)に示す)を文書ブロックの前後に追加しHTMLファイルを作成する(ステップ311)。

【0028】できたHTMLファイルは出力部16からコンピュータ23に表示される(ステップ312)。

【0029】図4はタグを元にしてキーワードを含む文書ブロックをHTMLファイルから抜き出す処理を示す流れ図である。

【0030】あらかじめ設定しておいた文書の構造的な区切りで使用されることの多いタグがHTMLファイル中のキーワードの前にあるかを検索し、キーワードの前において最も近い位置にあるタグを探す(ステップ402)。次に同様に設定したタグがキーワードより後ろにあるかを検索しキーワードの後ろにおいて最も近い位置にあるタグを探す(ステップ403)。

【0031】次に、キーワードの前後で見つかったタグを含む部分を抜き出す(ステップ404)。

【0032】通常分割に利用するHTMLのタグは図6(a)に示すように、開始タグと終了タグが存在するが、抜き出した結果、開始、終了のいずれかが不足の場合と過剰の場合とが起るのでこれをチェックする(ステップ405)。文書ブロックの先頭に終了タグがある場合と、文書ブロックの終端に開始タグがある場合は、そのタグは過剰であるので取り去る。その他の場合で開始タグと終了タグが一致していない場合は、不足と見なしタグを追加する(ステップ406)。

【0033】図5は文書ブロックをつなぎ併せる処理のステップ310の一例を示す流れ図である。文書ブロックが複数ある場合には、これらを一つの文書ブロックにまとめる。すなわち、ある文書ブロックAにつづく次の文書ブロックBがあるかを調べ(ステップ502)、なければ(ステップ502のN枝)、この動作は終了するが、あった場合には(ステップ502のY枝)、この文書ブロックAの最後に<HR>タグを追加して(ステップ503)その後次の文書ブロックBを追加し(ステップ504)、残余の文書ブロックがなくなるまで<HR>タグで繋げゆき、一つの文書ブロックに纏めてゆく。

【0034】図7はHTMLファイルからキーワードを含む文書ブロックを図4に従って抜き出した結果の例を示している。

【0035】以上説明したように、本実施の形態のHTML記述情報抜粋表示システムは、ホームページの中から必要と考えられる部分を自動的に抜粋して一つに纏めて表示することができ、本当に必要な情報の掲載されたページを従来よりも迅速にかつ、効率的に探索することができる。

【0036】本実施の形態の説明ではWWWのホームページを例にとって行なったが、本発明はこれに限定されるものではなく、HTML記述言語による記述情報についての抜粋表示についても適用できることは自明である。

【0037】

【発明の効果】以上説明したように、本発明のHTML記述情報抜粋表示システムは、HTML記述言語による記述情報の中から必要と考えられる部分を自動的に抜粋して一つに纏めて表示することができ、本当に必要な情報の掲載されたページを従来よりも迅速にかつ、効率的に探索することができるという効果を有している。

【図面の簡単な説明】

【図1】本発明のHTML記述情報抜粋表示システムの一実施の形態を示すブロック図である。

【図2】本実施の形態のHTML記述情報抜粋表示システムにおける一実施例の構成を示すブロック図である。

【図3】本実施の形態のHTML記述情報抜粋表示システムの動作の一例を示す流れ図である。

【図4】本実施の形態のHTML記述情報抜粋表示システムの情報抽出の動作を示す詳細流れ図である。

【図5】本実施の形態のHTML記述情報抜粋表示システムの文書ブロックつなぎ動作の詳細流れ図である。

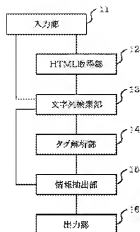
【図6】(a)は分割タグの一例を示すタグ図、(b)は必須タグの一例を示すタグ図である。

【図7】HTMLファイルからキーワードを含む文書ブロックを抜き出した結果の例を示す図である。

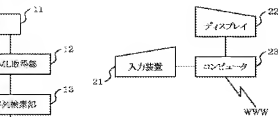
【符号の説明】

- 11 入力部
- 12 HTML取得部
- 13 文字列検索部
- 14 タグ解析部
- 15 情報抽出部
- 16 出力部
- 21 入力装置
- 22 デイスプレイ
- 23 コンピュータ

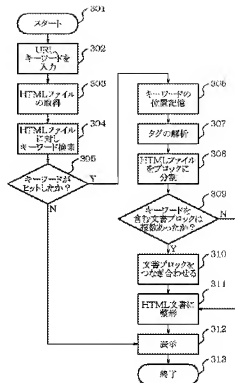
【図1】



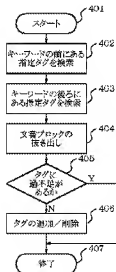
【図2】



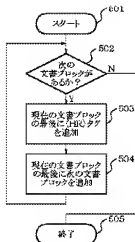
【図3】



【図4】



【図5】



【図6】

開始	終了	意味
0000Y	0100Y	書類の本文
0101	0201	見出し
0202	0302	見出し
0303	0403	見出し
0404	0504	見出し
0505	0605	見出し
0606	0706	見出し
0707	0807	表の作成
0808	0908	表の作成
0909	1009	表の作成
1010	1110	表の作成
1111	1211	表の作成
1212	1312	表の作成
1313	1413	表の作成
1414	1514	表の作成
1515	1615	表の作成
1616	1716	表の作成
1717	1817	表の作成
1818	1918	表の作成
1919	2019	表の作成
2020	2120	表の作成
2121	2221	表の作成
2222	2322	表の作成
2323	2423	表の作成
2424	2524	表の作成
2525	2625	表の作成
2626	2726	表の作成
2727	2827	表の作成
2828	2928	表の作成
2929	3029	表の作成
3030	3130	表の作成
3131	3231	表の作成
3232	3332	表の作成
3333	3433	表の作成
3434	3534	表の作成
3535	3635	表の作成
3636	3736	表の作成
3737	3837	表の作成
3838	3938	表の作成
3939	4039	表の作成
4040	4140	表の作成
4141	4241	表の作成
4242	4342	表の作成
4343	4443	表の作成
4444	4544	表の作成
4545	4645	表の作成
4646	4746	表の作成
4747	4847	表の作成
4848	4948	表の作成
4949	5049	表の作成
5050	5150	表の作成
5151	5251	表の作成
5252	5352	表の作成
5353	5453	表の作成
5454	5554	表の作成
5555	5655	表の作成
5656	5756	表の作成
5757	5857	表の作成
5858	5958	表の作成
5959	6059	表の作成
6060	6160	表の作成
6161	6261	表の作成
6262	6362	表の作成
6363	6463	表の作成
6464	6564	表の作成
6565	6665	表の作成
6666	6766	表の作成
6767	6867	表の作成
6868	6968	表の作成
6969	7069	表の作成
7070	7170	表の作成
7171	7271	表の作成
7272	7372	表の作成
7373	7473	表の作成
7474	7574	表の作成
7575	7675	表の作成
7676	7776	表の作成
7777	7877	表の作成
7878	7978	表の作成
7979	8079	表の作成
8080	8180	表の作成
8181	8281	表の作成
8282	8382	表の作成
8383	8483	表の作成
8484	8584	表の作成
8585	8685	表の作成
8686	8786	表の作成
8787	8887	表の作成
8888	8988	表の作成
8989	9089	表の作成
9090	9190	表の作成
9191	9291	表の作成
9292	9392	表の作成
9393	9493	表の作成
9494	9594	表の作成
9595	9695	表の作成
9696	9796	表の作成
9797	9897	表の作成
9898	9998	表の作成
9999	10000	表の作成

(a)

開始	終了	意味
0000Y	0100Y	書類の本文
0101	0201	見出し
0202	0302	見出し
0303	0403	見出し
0404	0504	見出し
0505	0605	見出し
0606	0706	見出し
0707	0807	表の作成
0808	0908	表の作成
0909	1009	表の作成
1010	1110	表の作成
1111	1211	表の作成
1212	1312	表の作成
1313	1413	表の作成
1414	1514	表の作成
1515	1615	表の作成
1616	1716	表の作成
1717	1817	表の作成
1818	1918	表の作成
1919	2019	表の作成
2020	2120	表の作成
2121	2221	表の作成
2222	2322	表の作成
2323	2423	表の作成
2424	2524	表の作成
2525	2625	表の作成
2626	2726	表の作成
2727	2827	表の作成
2828	2928	表の作成
2929	3029	表の作成
3030	3130	表の作成
3131	3231	表の作成
3232	3332	表の作成
3333	3433	表の作成
3434	3534	表の作成
3535	3635	表の作成
3636	3736	表の作成
3737	3837	表の作成
3838	3938	表の作成
3939	4039	表の作成
4040	4140	表の作成
4141	4241	表の作成
4242	4342	表の作成
4343	4443	表の作成
4444	4544	表の作成
4545	4645	表の作成
4646	4746	表の作成
4747	4847	表の作成
4848	4948	表の作成
4949	5049	表の作成
5050	5150	表の作成
5151	5251	表の作成
5252	5352	表の作成
5353	5453	表の作成
5454	5554	表の作成
5555	5655	表の作成
5656	5756	表の作成
5757	5857	表の作成
5858	5958	表の作成
5959	6059	表の作成
6060	6160	表の作成
6161	6261	表の作成
6262	6362	表の作成
6363	6463	表の作成
6464	6564	表の作成
6565	6665	表の作成
6666	6766	表の作成
6767	6867	表の作成
6868	6968	表の作成
6969	7069	表の作成
7070	7170	表の作成
7171	7271	表の作成
7272	7372	表の作成
7373	7473	表の作成
7474	7574	表の作成
7575	7675	表の作成
7676	7776	表の作成
7777	7877	表の作成
7878	7978	表の作成
7979	8079	表の作成
8080	8180	表の作成
8181	8281	表の作成
8282	8382	表の作成
8383	8483	表の作成
8484	8584	表の作成
8585	8685	表の作成
8686	8786	表の作成
8787	8887	表の作成
8888	8988	表の作成
8989	9089	表の作成
9090	9190	表の作成
9191	9291	表の作成
9292	9392	表の作成
9393	9493	表の作成
9494	9594	表の作成
9595	9695	表の作成
9696	9796	表の作成
9797	9897	表の作成
9898	9998	表の作成
9999	10000	表の作成

(b)

【図7】

